From Completing the Squares and Orthogonal Projection to Finite Element Methods

Mo MU

1 Background

In scientific computing, it is important to start with an appropriate model in order to design effective and efficient numerical methods for solving a given scientific problem. The modeling process usually involves approximations and assumptions, which leads to various mathematical formulations for different purposes. Different formulations may be mathematically equivalent in certain sense, but they may make significant differences numerically.

In the early stage of numerical PDEs, people started with the classical PDE form, say the Poisson equation for the elliptic model problem:

$$u_{xx} + u_{yy} = f, \qquad \text{in } \Omega \tag{1.1}$$

and applied finite differences to approximate the derivatives, which led to finite difference methods (FDM). The FDM approach has many difficulties, such as in handling complex geometry and boundary conditions, in convergence analysis, in strong regularity requirement, and in solving the resulting discrete system etc.

The Poisson equation can be derived from the variational principle in mechanics

$$\min_{u \in S} \left\{ \frac{1}{2} \int_\Omega (u_x^2 + u_y^2) - \int_\Omega fu \right\}, \tag{1.2}$$

where S is the feasible set of solutions. However, it is also difficult to derive and analyze numerical methods directly from the optimization formulation (1.2). In fact, there is another formulation between (1.1) and (1.2), the so called weak formulation, which is where the finite element methods (FEM) start with.

We will show how these different formulations are related to each other and how FEM can be derived and analyzed by simply using the ideas of

completing the squares and orthogonal projection. Through this example, we will demonstrate how these elementary, yet fundamental mathematical concepts and techniques are involved in advanced studies.

2 Optimization and Completing the Squares

2.1 Simplest case in $\mathbb{R}^1$

Consider the simplest optimization problem

$$\min_{x\in\mathbb{R}^1} f(x), \tag{2.1.1}$$

where

$$f(x) = \frac{1}{2}ax^2 - bx, \quad a > 0. \tag{2.1.2}$$

The most elementary approach to solve this problem is to reformulate it by completing the squares – just a little bit of algebra:

$$f(x) = \frac{1}{2}a(x - b/a)^2 - \frac{1}{2}b^2/a. \tag{2.1.3}$$

Thus, the optimization problem is equivalent to the equation

$$ax = b, \tag{2.1.4}$$

which, of course, can be solved immediately. Alternatively, one can use calculus to derive (2.1.4)by

$$f'(x) = 0. \tag{2.1.5}$$

2.2 Extension to $\mathbb{R}^n$

Consider

$$\min_{x\in\mathbb{R}^n} f(x), \tag{2.2.1}$$

where

$$f(x) = \frac{1}{2}x^T Ax - b^T x, \quad A_{n\times n} \text{ is symmetric and positive definite } ; x, b \in \mathbb{R}^n. \tag{2.2.2}$$

Now, completing the squares reads.

$$f(x) = \frac{1}{2}(x - A^{-1}b)^T A(x - A^{-1}b) - \frac{1}{2}b^T A^{-1}b. \tag{2.2.3}$$

Thus, the equivalent equation becomes

$$Ax = b. \tag{2.2.4}$$

Notice that (2.2.4) can also be written as

$$f'(x) = 0, \tag{2.2.5}$$

where f' denotes the Frechét derivative.

Observe that the framework in $\mathbb{R}^1$ is completely parallel to that in $\mathbb{R}^n$. Furthermore, $\mathbb{R}^n$ is the finite dimensional model of Hilbert spaces. So we can abstract and extend the framework to Hilbert spaces. For this purpose, let us introduce the bilinear form

$$a(u, v) = u^T Av, \quad \forall u, v \in \mathbb{R}^n, \tag{2.2.6}$$

and the linear functional

$$b(u) = b^T u, \quad \forall u \in \mathbb{R}^n. \tag{2.2.7}$$

Since A is symmetric and positive definite, $a(\cdot, \cdot)$ in fact defines an inner product in $\mathbb{R}^n$. The optimization problem can now be rewritten in terms of $a(\cdot, \cdot)$ and $b(\cdot)$. To cast equation (2.2.4) to a form in terms of $a(\cdot, \cdot)$ and $b(\cdot)$, we apply dot product on both sides of (2.2.4) with any $y \in \mathbb{R}^n$. It is easy to verify that (2.2.4) is equivalent to

$$a(x, y) = b(y), \quad \forall y \in \mathbb{R}^n. \tag{2.2.4$'$}$$

2.3 Extension to Hilbert Spare H

Given an inner product $a(\cdot, \cdot)$ and a continuous linear form b in H, consider the optimization problem

$$\min_{u \in H} f(u) \tag{2.3.1}$$

where

$$f(u) = \frac{1}{2}a(u, u) - b(u). \tag{2.3.2}$$

Completing the squares now reads

$$f(u) = \frac{1}{2}a(u-u^*, u-u^*) - \frac{1}{2}b(u^*), \tag{2.3.3}$$

where $u^* \in H$, such that

$$a(u^*, v) = b(v), \quad \forall v \in H. \tag{2.3.4}$$

The existence and uniqueness of the solution u^* of (2.3.4) is given by the well-known Lax-Milgram theorem. The proof of (2.3.3) is simply the same algebraic manipulation as used in completing the squares by expanding the terms and carrying out simplification:

$$\begin{aligned}
& \frac{1}{2}a(u-u^*, u-u^*) - \frac{1}{2}b(u^*) \\
= \; & \frac{1}{2}\left\{a(u,u) - a(u,u^*) - a(u^*,u) + a(u^*,u^\star)\right\} - \frac{1}{2}b(u^*) \\
= \; & \frac{1}{2}a(u,u) - a(u^*,u) + \left\{\frac{1}{2}a(u^*,u^\star) - \frac{1}{2}b(u^*)\right\} \\
= \; & \frac{1}{2}a(u,u) - b(u).
\end{aligned}$$

2.4 An example in $H^1(\Omega)$: Poisson equation

Consider the Sobolev space $H \equiv H_0^1(\Omega) = \{u | u \in L^2(\Omega), Du \in L^2(\Omega), u|_{\partial\Omega} = 0\}$.

$$a(u,v) = \int_\Omega u_x \cdot v_x + u_y \cdot v_y, \quad \forall u, v \in H,$$

and

$$b(v) = \int_\Omega f \cdot v, \quad f \in L^2(\Omega), \forall v \in H.$$

It can be shown that $a(\cdot,\cdot)$ and $b(\cdot)$ define an inner product (by Poincare's inequality) and a continuous linear form in H. So, the optimization problem (2.3.1) – (2.3.2) just corresponds to (1.2). The corresponding equivalent problem (2.3.4): find $u \in H_0^1(\Omega)$, such that

$$\int_\Omega u_x v_x + u_y v_y = \int_\Omega f \cdot v, \quad \forall v \in H_0^1(\Omega)$$

is the so-called weak formulation, which leads to the classical Poisson equation with Dirichlet boundary condition:

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases}$$

if $u \in H^2(\Omega)$ by integration by parts.

3 Orthogonal projection and FEM

3.1 Galerkin approximation

Starting with the abstract weak formulation P: find $u \in V$, such that

$$a(u, v) = f(v), \quad \forall v \in V, \tag{3.1.1}$$

where V is a Hilbert space, $a(\cdot, \cdot)$ is an inner-product in V, and $f(\cdot)$ is a continuous linear form in V. For the computational purpose, one needs to discretize the continuous problem P to a finite dimensional problem. Suppose we have a subspace of dimension N: $V_h \subset V$, the Galerkin approximation restricts P to V_h, which leads to a finite dimensional problem P_h: find $u_h \in V_h$, such that

$$a(u_h, v_h) = f(v_h), \quad \forall v_h \in V_h. \tag{3.1.2}$$

(3.1.2) can be represented as a linear system equation in terms of the N basis functions of V_h, where the coefficient matrix is also SPD because $a(\cdot, \cdot)$ is an inner product. Intuitively, if $V_h \to V$, say by increasing the dimension N of V_h, one could expect the convergence of P_h to P in certain sense. If this true, furthermore, is P_h the optimal approximation to P? The following "obvious" observation is crucial for understanding the convergence mechanism of this approach. From (3.1.1) and (3.1.2), we see that

$$a(u - u_h, v_h) = 0, \quad \forall v_h \in V_h. \tag{3.1.3}$$

The geometric interpretation is that the approximation u_h to u is nothing else, but the orthogonal projection of u to the subspace V_h with respect to the inner-production $a(\cdot, \cdot)$. Therefore, u_h is the optimal approximation of u in this sense because

$$\|u - u_h\|_a = \min_{v_h \in V_h} \|u - v_h\|_a \tag{3.1.4}$$

is the shortest distance, where $\|u\|_a = a^{1/2}(u,u)$ is the energy norm . This basic geometric concept leads to the fundamental of FEM convergence analysis.

3.2 FEM

Now the questions are: how to construct V_h, whether V_h converges to V and how u_h converges to u. To construct V_h, one first discretizes Ω by a mesh Ω_h, where h is the average spacing of the elements. Then, one constructs piecewise polynomials to form V_h. In the simplest case where continuous piecewise linear polynomials are used, which are called Courant elements, it can be shown that V_h is a subspace of $H^1(\Omega)$. Therefore, the Galerkin framework can be applied to such a V_h, which leads to the famous finite element methods.

3.3 Convergence analysis of FEM and interpolation theory in Sobolev spaces

The orthogonal projection property

$$\|u - u_h\|_a = \min_{v_h \in V_h} \|u - v_h\|_a$$

does not tell too much about how small $\|u - u_h\|_a$ is. If we can find a particular $\overline{u}_h \in V_h$ such that $\|u - \overline{u}_h\|_a$ can be estimated, then we have

$$\begin{aligned} \|u - u_h\|_a &= \min_{v \in V_h} \|u - v_h\|_a \\ &\leq \|u - \overline{u}_h\|_a. \end{aligned} \tag{3.3.1}$$

It can be shown that the energy norm is an equivalent norm of $\|\cdot\|_{H^1(\Omega)}$. Thus there exists a constant C, such that

$$\|u - u_h\|_{H^1(\Omega)} \leq C\|u - \overline{u}_h\|_{H^1(\Omega)}. \tag{3.3.2}$$

The natural choice of $\overline{u}_h$ is the interpolation of u, denoted by $I_h u$. Unfortunately, $H^1(\Omega)$ is not embedded in $C^0(\Omega)$. So for $u \in H^1(\Omega), I_h u$ is not defined. Notice that $H^2(\Omega)$ is embedded in $C^0(\Omega)$. Therefore, under the assumption on the smoothness of u: $u \in H^2(\Omega)$, we have

$$\begin{aligned} \|u - u_h\|_{H^1(\Omega)} &\leq C\|u - I_h u\|_{H^1(\Omega)} \\ &= C\|(I - I_h)u\|_{H^1(\Omega)}, \end{aligned} \tag{3.3.3}$$

which is the so-called Cea lemma. Thus the convergence analysis of FEM amounts to the approximation theory of interpolation in Sobolev spaces, which opened a new field in numerical analysis in 1960's. From this approximation theory, it holds, if $u \in H^2(\Omega)$,

$$\|u - I_h u\|_{H^1(\Omega)} \leq C\|u\|_{H^2(\Omega)} h. \tag{3.3.4}$$

Thus,

$$\|u - u_h\|_{H^1(\Omega)} \leq C\|u\|_{H^2(\Omega)} \cdot h. \tag{3.3.5}$$

Namely, $\|u - u_h\|_{H^1(\Omega)} \to 0$ as $h \to 0$ with the order of $O(h)$.

There are other computational issues that we omit here. In general, FEM overcomes many difficulties of FDM. The framework of FEM has been applied and extended to many applications, and has made enormously impact on scientific computing for several decades. It is elegant, yet the key ideas are easy to understand, which is the beauty of mathematics.

Reference P. G. Ciarlet, The Finite Element Method for Elliptic Problems, North-Holland Pub., 1978.